DOCUMENT RESUME

ED 078 072 TM 002 891

AUTHOR Gaines, W. George
TITLE Measuring Social Studies Achievement: Criterion-Referenced versus Norm-Referenced Tests for the Classroom Teacher.
PUB DATE 10 Nov 72
NOTE 23p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (New Orleans, Louisiana, November 10, 1972)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Academic Achievement; Academic Standards; *Criterion Referenced Tests; Educational Objectives; Evaluation Techniques; *Norm Referenced Tests; *Social Studies; *Test Construction

ABSTRACT

Criterion-referenced measurement has been hailed as one of the most significant developments in the recent history of educational evaluation. Its presence should be felt in the social studies over the next few years. The first part of the paper illustrates the differences between criterion-referenced and norm-referenced measurement--purpose for which the test is constructed; manner in which the test is constructed; specificity of information yielded; uses to be made of obtained test information--and traces the historical development of both, with particular attention to their influence on social studies testing. The second part of the paper describes the implementation of criterion-referenced testing in the social studies classroom--developing social studies objectives; setting appropriate standards; developing test items that measure attainment of objectives--and examines some of the major questions regarding this new trend in educational measurement. (Author)

FILMED FROM BEST AVAILABLE COPY

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED 078072

ED 078072 JUN 06 73

MEASURING SOCIAL STUDIES ACHIEVEMENT: CRITERION-REFERENCED VERSUS NORM-REFERENCED TESTS FOR THE CLASSROOM TEACHER

W. George Gaines
Louisiana State University
in New Orleans

Paper Presented at the Annual
Meeting of the Mid-South Educational
Research Association

New Orleans, Louisiana

November 10, 1972

MEASURING SOCIAL STUDIES ACHIEVEMENT: CRITERION-REFERENCED VERSUS NORM-REFERENCED TESTS FOR THE CLASSROOM TEACHER

Criterion-referenced measurement has been hailed as one of the most significant developments in the recent history of educational evaluation. Some experts predict that criterion-referenced tests will eventually fulfill many of the functions now being served by conventional forms of testing. At any rate, criterion-referenced measurement is one of the most visible topics in current measurement literature and should make its presence felt in the social studies over the next few years.

The first part of this paper will illustrate the differences between criterion-referenced and norm-referenced measurement and trace the historical development of both, with particular attention to their influence on social studies testing. The second part of the paper will describe the implementation of criterion-refernced testing in the social studies classroom and examine some of the major questions regarding this new trend in educational measurement.

Distinctions Between Criterion-Referenced and Norm-Referenced Measurement

As Popham and Husek (1971) noted, it is not possible to distinguish between a criterion-referenced test and a norm-referenced test simply by examining the items. This fact, however, does not diminish the importance of the distinctions between these two forms of testing. Several of the distinctions between criterion-referenced and norm-referenced tests

Abstract

Criterion-referenced measurement has been hailed as one of the most significant developments in the recent history of educational evaluation. Its presence should be felt in the social studies over the next few years. The first part of the paper illustrates the differences between criterion-referenced and norm-referenced measurement--purpose for which the test is constructed; manner in which the test is constructed; specificity of information yielded; uses to be made of obtained test information--and traces the historical development of both, with particular attention to their influence on social studies testing. The second part of the paper describes the implementation of criterion-referenced testing in the social studies classroom--developing social studies objectives; setting appropriate standards; developing test items that measure attainment of objectives--and examines some of the major questions regarding this new trend in educational measurement.

identified by Glaser and Nitko (1971) will serve as the basis of the discussion in this section of the paper.

Purpose for which the test is constructed. The basic purpose of a criterion-referenced test is to determine an individual's status with regard to some absolute standard. In other words, a criterion-referenced test is constructed to yield information that is interpretable in terms of some prespecified level of performance. The purpose of a norm-referenced test is to determine an individual's status with regard to the performance of others on the same test. In order to provide information about the relative standing of an individual in a group, norm-referenced tests are constructed by selecting those performances which are likely to maximize the opportunity to measure relative differences between individuals. Criterion-referenced tests tell us what a learner has achieved whereas norm-referenced tests tell us how much a learner has achieved with reference to his peers.

Criterion-referenced interpretations of test results are always between the score of an individual X_i and the criterion score X_c whereas norm-referenced test interpretations take into account the mean $\bar{X}.$ and standard deviation s_x of the group taking the test along with the individual's score X_i. The criterion score X_c is established prior to the test while the mean $\bar{X}.$ and standard deviation s_x are dependent upon the performance of every individual in the group taking the test.

Manner in which the test is constructed. A criterion-referenced test is constructed in two stages: first, by defining instructionally relevant

task domains and second, by selecting test items that are unbiased estimators of the learner's degree of competency in these domains. Norm-referenced tests are constructed in a different manner altogether. Although many norm-referenced tests are said to sample certain content areas, this type of sampling should not be equated with the notion of sampling from clearly defined and specific task domains. The selection of test items in the norm-referenced case is based upon statistical as well as content considerations. In order to have a test that produces the maximal number of discriminations between the individuals taking it, only items that approach the median difficulty level (half respond incorrectly) and correlate both positively and highly with total test score will be selected, other things being equal. These statistical considerations are irrelevant for item selection in criterion-referenced testing. There is reason to believe such practices are also damaging to the content validity of norm-referenced achievement tests (Anderson, 1972).

Specificity of information yielded. A criterion-referenced test is designed to yield information regarding an individual's competency in a specified task domain. To the extent that the task domain is clearly defined and the test items are representative samples of the task domain, the individual's performance on the test may be generalized to that task domain. In this case an individual's test score must unambiguously represent those tasks that he can perform.

Norm-referenced tests do not yield this type of information for two

reasons. First, norm-referenced test items are not constructed according to strict definitions of task domains; therefore, any match-up between a test item and a task domain of interest would be a fortuitous occurrence. Second, even though every test item on a norm-referenced test is conceivably a member of some task domain, the extent to which these items are representative of their domains is highly suspect due to their ex post facto identification. The hazards of making inferences based on an individual's response to norm-referenced test items has been acknowledged by norm-referenced test developers (Lindquist & Hieronymus, 1964) as well as their critics (Anderson, 1972).

Use to be made of the obtained test information. Both norm-referenced and criterion-referenced tests provide information which can be used for making decisions about individuals. It is the nature of the decision, however, that dictates which type of test should be used. For example, a social studies instructor who is teaching a unit on propaganda to a ninth-grade class needs to know whether each student can recognize the various propaganda techniques, say, in a daily newspaper editorial. A properly constructed criterion-referenced test would provide the instructor with information such as "Frank recognized instances of argumentum ad hominem in 90% of the cases but fails to exceed chance levels in recognizing argumentum ad authoritatem." A criterion-referenced test can provide an instructor with information that will help him decide what procedures can be followed to help Frank to master this skill. A norm-referenced test

in this situation would provide the instructor with information regarding Frank's relative position in the class. If Frank were above the class mean $\bar{X}$. on a norm-referenced test, the instructor might be led to the conclusion that "Frank is doing better than average and seems to know more about propaganda than Sue or Gary." On the other hand, if Frank's score X_F turned out to be far below the class mean $\bar{X}$. the instructor might be concerned yet unable to determine exactly where Frank was weak.

If an instructional system attempts to be adaptive to the learning needs of individuals, then its measurement needs will best be served by criterion-referenced tests. On the other hand, if an instructional system is selective as to learners it will admit and nonadaptive to individual learning needs, then it will need measures that spread individuals out on key ability dimensions--a task for which norm-referenced tests are best equipped.

Historical Development of Criterion-Referenced and Norm-Referenced Measurement in the Social Studies

Criterion-referenced measurement is not a new idea in education. As early as 1918, E. L. Thorndike noted the basic distinction between what we presently refer to as criterion-referenced and norm-referenced measurement. Criterion-referenced measurement played an important part in Washburne's (1922) Winnetka Plan and Morrison's (1926) work in mastery testing at the University of Chicago Experimental School.

One of the earliest uses of criterion-referenced testing in the social

studies was reported by Helen Boten (1932). Her contention was that classroom measurement in the social studies should be concerned with the development of tests that have diagnostic power for detecting faults in teaching and learning. Boten also maintained that these tests must be constructed around definite instructional objectives and the information provided by these tests be used for making decisions as to whether students should be given additional instruction or allowed to work on optional projects.

One of the most elaborate criterion-referenced test batteries of this period was constructed by Gladys Boyington (1932). Boyington's "Diagnostic Study Test" was designed to identify those students who had failed to master key social science concepts. This criterion-referenced test was accompanied by charts for monitoring individual performance according to problem types so that the teacher could plan both for individual and class remedial and advanced work.

In the social studies, as well as other areas of the curriculum, the ideas of mastery learning and criterion-referenced testing began to wane by the mid-thirties. Two primary reasons for this were (1) that measurement specialists were generally preoccupied with measuring trait variability and the relative differences between individuals and (2) subject matter specialists were generally reluctant to specify their goals in precise, observable learner behaviors. Both of these factors were readily apparent in the classic volume, Tests and Measurements in the Social Sciences, edited by T. L. Kelley and A. C. Krey (1934). Throughout the text of this record of the

four-year proceedings of the American Historical Association's Committee on Tests and Measurements, the clash between the thinking of specialists in measurement and social science content areas is painfully evident. The measurement specialists, headed by Kelley, tried to get the social science content specialists, headed by Krey, to formulate precise statements of the purposes and ends of social science instruction so that accurate measures might be developed. The social science content specialists contended that the goals of social science instruction were so complex as to defy specification. The two groups were virtually at an impasse' from this point on.

Another divisive issue was the manner in which test items were to be selected. The measurement specialists required that the final selection of test items be based on statistical considerations (i.e., difficulty level and discriminating power), criteria that the subject matter specialists believed was to blame for the inclusion of some mediocre items and the exclusion of items which they believed to be some of the very best. In one of the most spirited debates of the committee, Edith Parker, a content specialist, defended the choice of items on her geography test (which had extremely low reliabilities) by drawing an analogy between test items and a physician's thermometer. The thermometer is not discarded because the temperature of the patient is not normal; it is not the items which are poor but the people who were tested.

Tests and Measurements in the Social Sciences merits a special place in this review because it illustrates many of the measurement problems

in the social studies that are still with us in the seventies, namely, the reluctance of educators to specify the outcomes of instruction and our obcession with the norm-referenced measurement model.

The domination of norm-referenced measurement in the social studies can be seen in the writings of Anderson and Lindquist (1932), Wrightstone (1937), and Wilson and Murra (1938). The 1965 Yearbook of the National Council of the Social Studies, Evaluation in Social Studies, (Berg, 1965), was written almost exclusively from a norm-referenced measurement frame of reference. Throughout this long drought, Ralph Tyler (1938, 1967) insisted that norm-referenced tests are products of assumptions and conditions that are simply not applicable to our measurement needs in classrooms. Tyler and many others sharing similar views on the value of criterion-referenced testing have, in recent years, made a substantial impact in other areas of the curriculum, particularly in mathematics and reading. It would seem to be only a matter of time before criterion-referenced measurement "reappears" in the social studies.

Problems in Implementing Criterion-Referenced Testing Procedures in the Social Studies Classroom

Airasian and Madaus (1972) have identified three steps in implementing criterion-referenced testing procedures in the classroom: (1) to develop, prior to instruction, a list of objectives that identify the learner performances and products that are the desired outcomes of instruction; (2) to decide upon the nature of the standards which will be used to judge whether a learner's

performance or product indicates mastery of the instructional objective; and (3) to devise situations which allow the learner the opportunity to exhibit the desired performance or product. The remaining discussion will follow these headings.

Development of social studies objectives. The previous section on the historical background of measurement in the social studies alluded to the reluctance of educators to specify social studies objectives. Although resistance to instructional objectives among social studies educators appears to be declining, the general ambiguity of social studies objectives remains a persistent problem for measurement. Why are social studies objectives so ambiguous? One reason pointed out by Orlandi (1971) is that the controversial character of social studies has tended to make educators fearful of too much specificity and possible attacks from outside groups. Certainly another reason for the ambiguity is the honest and open disagreement among educators as to what the ultimate aims of social studies should be.

Questions as to the relative worth or significance of given objectives or goals are ultimately answerable on the basis of value preferences. This paper neither proposes nor advocates one set of objectives or goals over another. It is the position of this paper, however, that most social studies instruction is aimed at effecting some type of behavioral change in learners and that these intentions ought to be made explicit.

The argument that the goals of social studies instruction are so

profound and complex as to defy translation into observable terms may have served as a convenient buffer 40 years ago, but such a claim cannot be taken seriously today. Typically, the educators who object most vehemently to stating instructional objectives are those who have never been very successful at it. There is evidence that after educators are successfully trained in the techniques of writing objectives, they will tend to view instructional objectives as more valuable, more expressive, more powerful, and less threatening than before training (Jongsma & Gaines, 1972). Although there may be some reasonable objections to instructional objectives which stem from their misuse, the unambiguous specification of the tasks learners are expected to perform as a consequence of instruction is a prerequisite for criterion-referenced measurement.

Setting appropriate standards. The second step in implementing criterion-referenced measurement techniques in the social studies classroom is the setting of standards that will be used as a basis for inferring whether a learner's performance or product represents mastery of the instructional objectives. Glaser and Nitko (1971) have defined "mastery" in this context as meaning

> that an examinee makes a sufficient number of correct responses on the sample of test items presented to him in order to suggest the generalization (from this sample of items to the domain or universe of items implied by an instructional objective) that he has attained the desired, pre-specified degree of preficiency with respect to the domain. (Glaser & Nitko, 1971, p. 641)

In instances where the task domain consists of more than one test item, principles of hypothesis testing can be utilized in determining whether the learner has reached mastery of the objectives. In the special case when there is but one item in the task domain, a high confidence level might necessitate retesting.

Most social studies teachers probably do set some sort of standards for their students even though these standards are often implicit and individualistic. But if criterion-referenced measurement is to prove useful in the classroom, standards must be made explicit. Below are two sample objectives for a fifth-grade social studies class.

1. Given a list of the 50 states, the student will write the name of the capital beside the state with 90% accuracy.
2. Given a list of 10 quotations from a previously witnessed role-play on the theme of ethnocentrism, the student will identify those quotations which reflect ethnocentric attitudes on the part of the speaker with 90% accuracy.

The procedures for assessment outlined above can help us determine in a rather straightforward manner whether or not a learner has "mastered" each of these objectives; such procedures cannot, however, help us determine which objective is the more important.

One way of assessing the importance of an instructional objective is in terms of its ultimate benefit to the learner; another way is in terms of the degree to which other learnings depend on that objective. The past

few years have seen efforts to determine the ordering of or sequential relationships between learnings in a content area. (Airasian, 1971; Gagne', 1965; Resnick, 1967). Such an approach is known as task analysis. A task analysis in social studies would involve the breaking down of a general objective into its various tasks and subtasks. Through task analysis it may be possible to identify certain learnings that are prerequisite to a large number of learnings. If this were the case the specific objectives for the "prerequisites" would assume more instructional importance than specific objectives considered ancillary to the general objective. Do such hierarchies exist in social studies? Too little work has been done in the area to say for certain although the loosely organized content of the social studies makes one think that task analysis would not prove as productive as it already has in the more organized disciplines of mathematics and chemistry.

The importance of an instructional objective needs to be assessed, both in terms of its ultimate worth to the learner and the extent to which other learnings may be dependent upon it, because the objective's importance has fundamental implications for the standard of mastery to be employed. For example, if objective number one above were viewed as unimportant in relation to other fifth-grade social studies objectives, we might want to lower the standard of mastery or perhaps restrict it to a selected group of states.

These and other considerations have been taken into account in

Emrick's (1971) mastery test model. Specifically, the importance of an objective is included in what Emrick terms the "ratio of regret." The model is summarized by an algorithm for determining the optimal cut-off score on a mastery test. This score is derived from estimates of the relative decision error costs and relative item error probabilities associated with the test. Although Emrick's model may ultimately lead to setting appropriate standards for criterion-referenced tests, his method for determining the ratio of regret appears somewhat arbitrary and in need of further elaboration.

Developing test items that measure attainment of objectives. For a test to have a high degree of content validity it must be demonstrated that the test items adequately represent the task domains specified by the test objectives (American Psychological Association, 1966). Content validity, of course, is of the utmost importance to a criterion-referenced test.

How is content validity determined? The generally recommended technique is to use one's judgment of the extent to which given test items appear to be related to the objectives in question. According to Cronbach (1971), the only requirement is that the boundaries of the task domain be clearly defined so that reasonable observers can agree on which items are included by the definition and which are not. While this procedure may be appropriate for some norm-referenced tests, there is reason to believe that it is not rigorous enough for criterion-referenced tests.

Tyler (1967) warned us that "little theory has been formulated or techniques devised to aid in the construction of relatively homogeneous

samples of exercises faithfully reflecting an educational objective" (page 14). Since the time of his writing a number of techniques have been proposed for developing test items that are "faithful reflections" of instructional objectives. These proposals range from rather simple to rather complex, and it is not within the scope of this paper to enumerate or discuss them in any great detail. Instead, examples will be provided that serve to illustrate the kinds of thinking going on in this area.

The generation of test items for a given objective will be a function of the level of specificity of that objective. A narrowly defined task domain will generate fewer test items than a broadly defined domain, other things being equal. Klein (1970) has suggested that objectives be written "at a level of generality that will be interpretable to the person who has to use the test results" (page 3). While Klein's statement could hardly be termed "operational," his use of sample objectives and matching test items suggests that the kinds of objectives he prefers are those that imply a relatively large task domain. Klein's examples are drawn from mathematics, and since this paper is addressed to an audience of social studies educators, the writer will run the risk of translating Klein's mathematics examples into social studies examples.

Sample objective: The student will be able to identify artifacts.

Sample items: 1. Which is an artifact?

a. cow

b. boy

c. hat

d. hill

2. "Go to the bulletin board and point to the artifacts in the picture."

3. "Look at the objects on the table. Sort them into two groups: those that are artifacts and those that are not artifacts."

4. "I'm going to call out some names of things. Tell me if the thing is an artifact."

These sample test items were included to illustrate four of the possible item forms that can be used in assessing mastery of this objective. Manipulation of the difficulty of the task is possible by varying the stimuli the learner must examine; for example, if the foils in the first item form changed to "kimono," "etouffe'," "puma," and "amoeba." The point is that all of these item forms and item difficulties are "faithful representatives" of the stated objective. Klein, however, sees no problem with general objectives. His recommendation is that the test constructor should try to cover both the range of formats and difficulties on the test. In many cases, using Klein's technique, not only would the sampling problems be almost insurmountable but important information about what the student could do would be lost. And what about content validity? Adherence to this technique would satisfy Cronbach's definition, but would it be safe to make inferences about such a large task domain, given a small and perhaps biased sample of test items?

The other end of the complexity continuum is represented by the

work of Hively (1966) and his associates (Hively, Patterson, & Page, 1968). Essentially their approach is one of stating a general task format accompanied by a list of rules for generating every possible test item in a task domain. With all possible items for a task domain having been identified, content validity decisions would then be quite clear cut.

Another approach to developing test items based on instructional objectives is to follow the "instructional specification" outline proposed by Sullivan, Baker, and Schutz (1971).

The instructional specification (I.S.) requires that an instructional objective state the behavior a learner is to exhibit and the given conditions under which the behavior is to be demonstrated. Mastery test items must match the specifications of the objective exactly. Transforming our previously illustrated social studies objective so that it meets these criteria, we find that it must be narrowed in scope.

Objective: (revised)	The learner will identify examples of artifacts, given a set of objects that includes both artifacts and non-artifacts.
Test item:	"Look at the objects on the table. Sort them into two groups: those that are artifacts and those that are not artifacts."

A second important component of the I.S. is that of "limits" or characteristics of the correct responses and of the plausible, but incorrect responses for the desired behavior. If we assess the learner only by using examples that are obvious, such as an apple, a toy car, a red rose, an empty glass, and so forth, we may develop the illusion that the learner

is proficient when in reality he is not. It stands to reason that the examples (both positive and negative) of the concept "artifact" should be those which have the greatest probability of eliciting incorrect responses (e.g., wooden carving, artificial apple, cherry pie, bird's nest, etc.). While it is not essential, "limits" could be built into the given conditions component of an objective.

Sullivan, Baker, and Schutz also indicate that the examples used in assessment should be different from those provided during instruction. Anderson (1972) and Jenkins and Deno (1971) have also presented arguments in support of this point. If, in testing for comprehension of a concept, the substantive language of instruction and the test are not different, the examinee may make the correct response solely as a result of orthographic or phonological overlap.

In summary, the development of test items that measure attainment of instructional objectives is subject to at least four constraints:

1. The test item must elicit the behavior called for in the objective.
2. The test item must involve the exact given conditions denoted in the objective.
3. The limits of the correct and incorrect responses or response choices of the test item must be clearly defined.
4. The examples in the test item must be different from those used during instruction.

These restraints represent at least a prototype set of rules for assessing the content validity of criterion-referenced measures. An appealing

attribute of these rules is that their necessity can be established empirically. We (Gaines & Jongsma, 1972) decided to violate rule number two, "given conditions," in assessing first-graders' mastery of our revised objective on identifying artifacts. Since the stimulus called for in the objective was "objects," we deliberately used two additional stimuli, drawings of the objects and oral words for the objects, to see if this made any difference. Apparently it did because pupils who were administered the oral words format scored significantly lower than pupils tested on the drawings and objects formats. No difference was found between the drawings and objects formats. These findings suggest that major deviations from the "given conditions" rule can change the nature of the task to the extent that the test items associated with the deviations can not be considered members of the task domain implied by the objective.

These prototype rules invite the attention of researchers.

REFERENCES

Airasian, P. W. A study of the behaviorally dependent, classroom taught task hierarchies. Educational Technology Research Report Series, Number 3, 1971.

Airasian, P. W. & Madaus, G. F. Criterion-referenced testing in the classroom. Measurement in Education, National Council on Measurement in Education. 1972, 1 (3).

American Psychological Association, Standards for educational and psychological tests and manuals. Washington: APA, 1966.

Anderson, R. C. How to construct achievement tests to assess comprehension. Review of Educational Research, 1972, 42 (2).

Berg, H. D. (ed). Evaluation in social studies, Thirty-fifth Yearbook of the National Council for the Social Studies. Washington: National Council for the Social Studies, 1965.

Boten, H. A testing-teaching scheme for senior high school American history. In Classroom and administrative problems in the teaching of the social sciences, Second Yearbook of the National Council of the Social Studies. Philadelphia: McKinley Publishing Company, 1932, 208-212.

Boyington, G. Experiments with diagnostic tests to determine knowledge of study tools and techniques in the social studies. In Classroom and administrative problems in the teaching of the social sciences. Second Yearbook of the National Council of the Social Studies. Philadelphia: McKinley Publishing Company, 1932, 132-163.

Cronbach, L. J. Test validation. In Thorndike, R. L. (Ed.) Educational measurement. Washington: American Council on Education, 1971, 443-507.

Emrick, J. A. An evaluation model for mastery testing. Journal of Educational Measurement, 1971, 8(4), 321-326.

Gagne, R. M. The conditions of learning. New York: Holt, Rinehart & Winston, 1965.

Gaines, W. G. & Jongsma, E. A. The effects of mode of test stimuli on the performance level of first-grade pupils. In press, 1972.

Glaser, R., Nitko, A. J., & Thorndike, R. L. (Ed.). Measurement in learning and instruction. In Educational Measurement. Washington: American Council on Education, 1971, 625-670.

Hively, W. Preparation of a programmed course in algebra for secondary school teachers: A report to the National Science Foundation. Minnesota State Department of Education, Minnesota National Laboratory, 1966.

Hively, W., Patterson, H. L., & Page, S. A "universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5, 275-290.

Jenkins, J. R. & Deno, S. L. Assessing knowledge of concepts and principles. Journal of Educational Measurement, 1971, 8(2), 95-101.

Jongsma, E. A. & Gaines, W. C. The effectiveness of an in-service training session dealing with instructional objectives. In press, 1972.

Kelley, T. L. & Krey, A. C. Tests and measurements in the social sciences. Part IV of the Report of the commission on the social studies. New York: Charles Scribner's Sons, 1934.

Klein, S. Evaluating tests in terms of the information they provide. Evaluation Comment, 1969, 2(2), 3-4.

Lindquist, E. F., & Hieronymus, A. N. Teachers manual: Iowa tests of basic skills. Boston: Houghton-Mifflin, 1964.

Morrison, H. C. The practice of teaching in the secondary school. Chicago: University of Chicago Press, 1926.

Orlandi, L. R. Evaluation of learning in secondary school social studies. In Bloom, B. S., Hastings, J. T. & Madaus, G. F. (Eds.), Handbook on formative and summative evaluation of student learning. New York: McGraw-Hill, 1971, 449-498.

Popham, W. J. and Husek, T. R. Implications of criterion-referenced measurement. In Popham, W. J. (Ed.), Criterion-referenced measurement. Englewood Cliffs, N. J.: Educational Technology Publishers, 1971, 17-37.

Resnick, L. B. Design of an early learning curriculum. University of Pittsburgh Learning Research and Development Center. Working paper 16, 1967.

Sullivan, H. J., Baker, R. L. & Schutz, R. E. Developing Instructional Specifications. In R. L. Baker & R. E. Schutz (Eds.), Instructional Product Development. New York: Van Nostrand Reinhold Company, 1971, 66-98.

Thorndike, E. L. The nature, purposes and general methods of measurements of educational products. In Whipple, G. M. (Ed.), The measurement of educational products. Seventeenth Yearbook of the National Society for the Study of Education, Part II. Bloomington Public School Publishing Co., 1918, 16-24.

Tyler, R. W. The specific techniques of investigation: examining and testing acquired knowledge, skill, and ability. In The scientific movement in education, The Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II. Bloomington, Illinois: Public School Publishing Company, 1938, 341-355.

Tyler, R. W., Changing concepts of educational evaluation. In Stake, R. (Ed.), Perspectives of curriculum evaluation. American Educational Research Association Monograph Series on Curriculum Evaluation, Chicago: Rand McNally, 1967, 13-18.

Washburne, C. W. Educational measurements as a key to individualizing instruction and promotions. Journal of Educational Research, 5, 1922 195-206.

Wilson, H. E. & Murra, W. F. Contributions of research to special methods: the social studies. In The scientific movement in education, The Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II. Bloomington, Illinois: Public School Publishing Company, 1938, 147-160.

Wrightstone, J. W. Testing in the social studies. In Barnes, C. C. (Ed.) The contribution of research to the teaching of the social studies, National Council for the Social Studies Eighth Yearbook. Cambridge: National Council for the Social Studies, 1937, 207-239.